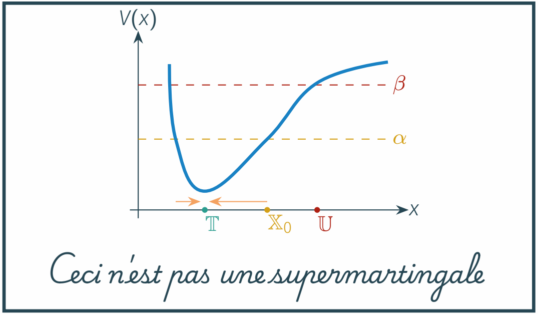
The Treachery of Certificates



Greg Neustroev • Zeroth Research • June 25, 2026 (Topos Institute, Oxford Seminar)

The problem

The task: Prove a behavioral property

- Think of a *navigating robot*.

The task: Prove a behavioral property

- Think of a *navigating robot*.
- It operates in *continuous time*.

The task: Prove a behavioral property

- Think of a *navigating robot*.
- It operates in *continuous time*.
- It is affected by exogenous *noise*.

The task: Prove a behavioral property

- Think of a *navigating robot*.
- It operates in *continuous time*.
- It is affected by exogenous *noise*.
- We want it to be *AI-controlled*.

The task: Prove a behavioral property

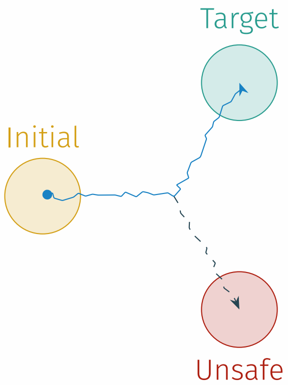
- Think of a *navigating robot*.
- It operates in *continuous time*.
- It is affected by exogenous *noise*.
- We want it to be *AI-controlled*.
- To do so safely, we want some *quanititative guarantees*, such as the probability that it reaches the target $\mathbb{T}$ without becoming unsafe $\mathbb{U}$:

$$\mathbb{P}\left[\text{reach } \mathbb{T} \text{ before } \mathbb{U}\right] \geq 1 - \delta.$$

The task: Prove a behavioral property

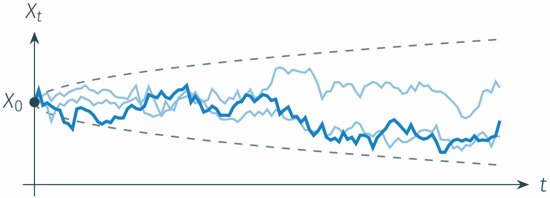
- Think of a *navigating robot*.
- It operates in *continuous time*.
- It is affected by exogenous *noise*.
- We want it to be *AI-controlled*.
- To do so safely, we want some *quanititative guarantees*, such as the probability that it reaches the target $\mathbb{T}$ without becoming unsafe $\mathbb{U}$:

$$\mathbb{P}\big[\text{reach } \mathbb{T} \text{ before } \mathbb{U}\big] \geq 1 - \delta.$$



The model: A stochastic differential equation

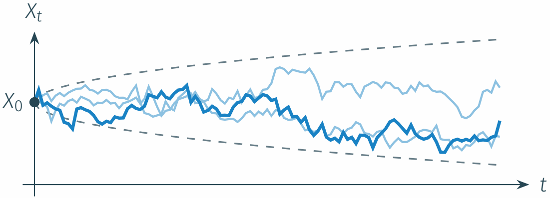
The state X_t wanders in *continuous time*



The model: A stochastic differential equation

The state X_t wanders in *continuous time* under *drift* and *diffusion*:

$$\mathrm{d}X_t \;=\; \underbrace{f(X_t)}_{\text{drift}}\,\mathrm{d}t \;+\; \underbrace{g(X_t)}_{\text{diffusion}}\;\mathrm{d}W_t.$$



Why it is hard (and the idea)

- We cannot simulate our way to a guarantee: the horizon is *infinite* and we would have to integrate over *all of the randomness*.

Why it is hard (and the idea)

- We cannot simulate our way to a guarantee: the horizon is *infinite* and we would have to integrate over *all of the randomness*.
- Instead, exhibit a single function on states whose values *prove* the guarantee — no unrolling of the dynamics.

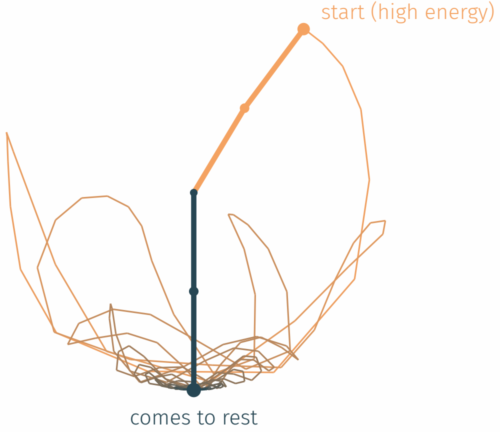
Why it is hard (and the idea)

- We cannot simulate our way to a guarantee: the horizon is *infinite* and we would have to integrate over *all of the randomness*.
- Instead, exhibit a **single function on states** whose values *prove* the guarantee — no unrolling of the dynamics.

That function is the supermartingale certificate.

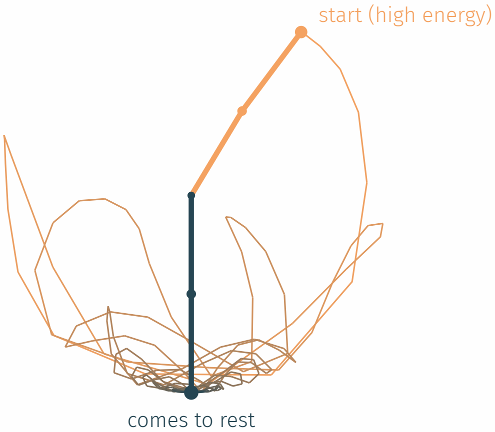
The rest of the talk: what it is, how we find it, and what it is *not*.

An example: Double pendulum



Prove to me that the pendulum with friction eventually stops — without solving the motion.

An example: Double pendulum



Prove to me that the pendulum with friction eventually stops — without solving the motion.

Energy constantly dissipates, therefore it will reach zero. *Energy is the certificate*.

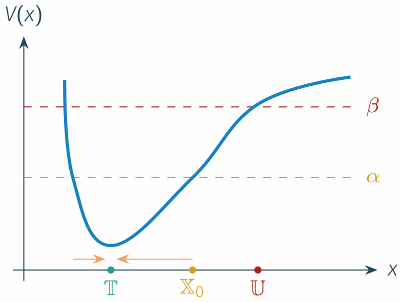
Exhibit A: The certificate — a function we synthesise

Exhibit A: The certificate

A twice-continuously differentiable function $V : \mathbb{X} \to \mathbb{R}_{\geq 0}$ obeying *pointwise* inequalities:

- $V \geq 0$ everywhere;
- $V \leq \alpha$ on the initial set $\mathbb{X}_0$;
- $V \geq \beta$ on the unsafe set $\mathbb{U}$ (with $\beta > \alpha$);
- $[\mathrm{G}V](x) \leq -\varepsilon$ for $x \notin \mathbb{T} \cup \mathbb{U}$ (decreases on average).

G is the *infinitesimal generator*: the expected *rate of change* of V along the noisy trajectory.



$$\mathbb{P}\left[\text{reach } \mathbb{T} \text{ before } \mathbb{U}\right] \geq 1 - \frac{\alpha}{\beta}.$$

Finding it is the hard part

good news

Checking a candidate V is easy: verify the inequalities — locally, symbolically, once and for all.

bad news

Finding a good V is a search over an *enormous space of functions* — the real bottleneck.

Verification needs nothing fancy. The search does.

Neural certificates: Learn the scalar

The requirements are just *inequalities*:

$$V(x) \leq \alpha \text{ on initial states } \mathbb{X}_0$$

$$V(x) \geq \beta \text{ on unsafe states } \mathbb{U}$$

$$[GV](x) \leq -\varepsilon \text{ on decrease domain}$$

- Any function meeting them is a certificate.

Neural certificates: Learn the scalar

The requirements are just *inequalities*:

$$V(x) \leq \alpha \text{ on initial states } \mathbb{X}_0$$
$$V(x) \geq \beta \text{ on unsafe states } \mathbb{U}$$
$$[GV](x) \leq -\varepsilon \text{ on decrease domain}$$

- Any function meeting them is a certificate.
- A neural network is a function.

Neural certificates: Learn the scalar

The requirements are just *inequalities*:

$$V(x) \leq \alpha \text{ on initial states } \mathbb{X}_0$$
$$V(x) \geq \beta \text{ on unsafe states } \mathbb{U}$$
$$[GV](x) \leq -\varepsilon \text{ on decrease domain}$$

- Any function meeting them is a certificate.
- A neural network is a function.
- So: train V_θ to *minimize the violation*.

Neural certificates: Learn the scalar

The requirements are just *inequalities*:

$$V(x) \leq \alpha \text{ on initial states } \mathbb{X}_0$$

$$V(x) \geq \beta \text{ on unsafe states } \mathbb{U}$$

$$[\mathrm{G}V](x) \leq -\varepsilon \text{ on decrease domain}$$

- Any function meeting them is a certificate.
- A neural network is a function.
- So: train V_θ to *minimize the violation*.

Turn inequalities into losses

e.g. for $V_\theta(x) \geq \beta$ on $\mathbb{U}$:

$$\min_\theta \sum_{x \in \text{batch} \cap \mathbb{U}} (\beta - V_\theta(x))^+$$

Push every violation down to zero in this way.

Exhibit B: The supermartingale — a process that proves

To speak of a *stochastic process*: Three ingredients

All intuitive:

- the random trajectory $(X_t)_{t\geq 0}$ — one continuous path per outcome;

To speak of a *stochastic process*: Three ingredients

All intuitive:

- the random trajectory $(X_t)_{t\geq 0}$ — one continuous path per outcome;
- the filtration $(\mathcal{F}_t)_{t\geq 0}$ — "what is known by time t", information that *grows*;

The measure-theoretic rigour (filtrations, integrability) is real and available — but the intuition above is the whole content.

To speak of a *stochastic process*: Three ingredients

All intuitive:

- the random trajectory $(X_t)_{t\geq 0}$ — one continuous path per outcome;
- the filtration $(\mathcal{F}_t)_{t\geq 0}$ — "what is known by time t", information that *grows*;
- the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}_t]$ — the *best forecast* of the future given the history.

The measure-theoretic rigour (filtrations, integrability) is real and available — but the intuition above is the whole content.

What is a supermartingale?

The gambling picture

A process $(M_t)_{t\geq 0}$, adapted to $(\mathcal{F}_t)$, is a *supermartingale* if

$$\mathbb{E}\big[M_t \mid \mathcal{F}_s\big] \;\leq\; M_s \qquad \text{for all } s \leq t.$$

"Given everything up to time s, my expected fortune at any later t is *no larger*."

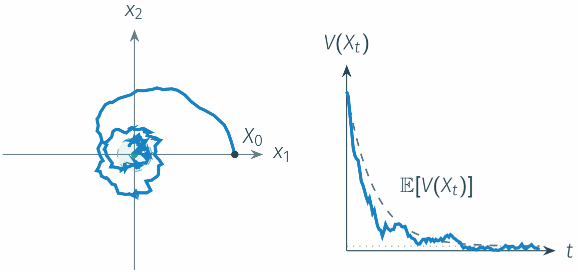
- *martingale*: equality (a fair game);
- *super*martingale: $\leq$ (unfavourable).

Bounded below *and* non-increasing on average $\Rightarrow$ it must *settle*. That is the engine of every proof here.

Let the function *ride* the trajectory

A concrete system — a *stable spiral* (eigenvalues $-a \pm i\omega$, $a > 0$) — with the simplest certificate $V(x) = \|x\|^2$:

$$\mathrm{d}X_t = A\,X_t\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \quad A = \begin{pmatrix} -a & -\omega \\ \omega & -a \end{pmatrix}; \qquad [GV](x) = -2a\,V(x) + 2\sigma^2.$$

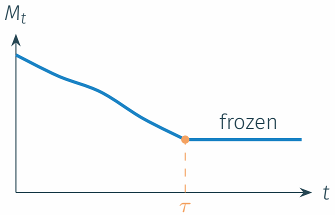


The proof rides on the *stopped* process

The drift $GV \leq -\varepsilon$ holds *only off* $\mathbb{T} \cup \mathbb{U}$. So the supermartingale is not $V(X_t)$ itself, but the stopped process — with τ read off a *level set of V*:

$$\tau = \inf\{\, t : V(X_t) \geq \beta \;\text{ or }\; X_t \in \mathbb{T} \,\}, \qquad M_t = V\big(X_{t \wedge \tau}\big).$$

Up to τ the state is off $\mathbb{T} \cup \mathbb{U}$ (there $\{V < \beta\}$ misses $\mathbb{U}$), so (M_t) is a *nonnegative supermartingale* — constant after τ.



The one theorem we need: Optional stopping

Optional stopping

Let (M_t) be a *nonnegative* supermartingale and τ a stopping time with $\tau < \infty$ almost surely. Then

$$\mathbb{E}[M_\tau] \ \leq \ \mathbb{E}[M_0].$$

This is the bridge from *pointwise inequalities* to a *probability bound*.

Note the side-condition (nonnegativity) — keep an eye on it: it is exactly the kind of extra data the bare function does *not* carry.

The ε-drift buys finiteness (reach)

Dynkin's formula

For a stopping time τ and $t \geq 0$,

$$\mathbb{E}\big[V(X_{t\wedge\tau})\big] \;=\; V(x_0) + \mathbb{E}\Big[\int_0^{t\wedge\tau} [GV](X_s)\,\mathrm{d}s\Big].$$

Up to τ the state is off $\mathbb{T} \cup \mathbb{U}$, so $GV \leq -\varepsilon$ there and the integral is $\leq -\varepsilon\,\mathbb{E}[t \wedge \tau]$.
With $V \geq 0$, letting $t \to \infty$ (monotone convergence),

$$\varepsilon\,\mathbb{E}[\tau] \;\leq\; V(x_0) \qquad \Longrightarrow \qquad \mathbb{E}[\tau] \;\leq\; V(x_0)/\varepsilon \;<\; \infty.$$

The process stops almost surely in finite expected time — so $\tau < \infty$ a.s., the hypothesis optional stopping needs.

Assemble the reach–avoid bound (avoid)

Ville's maximal inequality

For a *nonnegative* supermartingale (M_t) and any level $\lambda > 0$,

$$\mathbb{P}\left[\sup_{t\geq 0} M_t \;\geq\; \lambda\right] \;\leq\; \mathbb{E}[M_0]/\lambda.$$

Apply it to $M_t = V(X_{t\wedge\tau})$ at level β — controlling $\mathbb{U}$ through a *level set of V itself*:

$$\mathbb{P}\left[\sup_{t\geq 0}\; V(X_{t\wedge\tau}) \;\geq\; \beta\right] \;\leq\; V(x_0)/\beta.$$

Since $\mathbb{U} \subseteq \{V \geq \beta\}$, entering $\mathbb{U}$ forces the level to be crossed, so

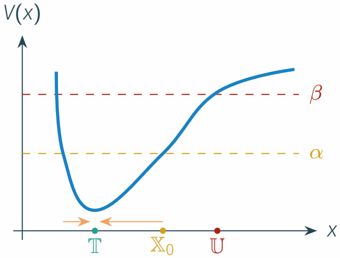
$$\mathbb{P}[\text{reach } \mathbb{T},\ \text{avoid } \mathbb{U}] \;\geq\; 1 - V(x_0)/\beta \;\geq\; 1 - \frac{\alpha}{\beta}$$

From certificate to supermartingale

A function is less data than a process

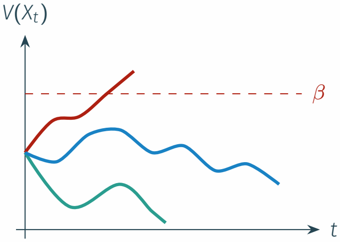
$$V : \mathbb{X} \to \mathbb{R}_{\geq 0}$$

Exhibit A — a *function*, one number per state



$$(M_t)_{t\geq 0} = (V(X_t))_{t\geq 0}$$

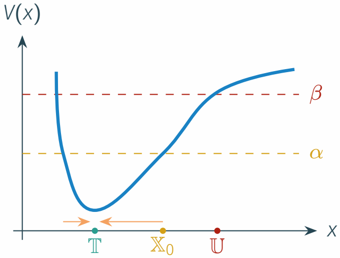
Exhibit B — a *process* over time and randomness



A function is less data than a process

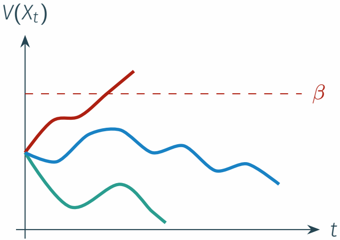
$$V : \mathbb{X} \to \mathbb{R}_{\geq 0}$$

Exhibit A — a *function*, one number per state



$$(M_t)_{t\geq 0} = (V(X_t))_{t\geq 0}$$

Exhibit B — a *process* over time and randomness



V is everything on the left. To even *write down* M_t you must add more things: the *dynamics* (how X moves), the *filtration* $(\mathcal{F}_t)$ (what is known when), and a *stopping rule* τ. The supermartingale certificate function V is one input among four.

Why you must stop at all

The naive process fails.

The raw $N_t = V(X_t)$ need not be a supermartingale: the certificate buys decrease *only off the sets*,

$$\mathbb{E}[V(X_{t+1}) \mid X_t{=}x] \le V(x) - \varepsilon, \quad x \notin \mathbb{T} \cup \mathbb{U}.$$

On $\mathbb{T} \cup \mathbb{U}$ nothing constrains the drift — after reaching $\mathbb{T}$ the dynamics can carry V back *up*. So $\mathbb{E}[N_{t+1} \mid \mathcal{F}_t] \le N_t$ can fail, and the proof rule is simply inapplicable.

Why you must stop at all

The naive process fails.

The raw $N_t = V(X_t)$ need not be a supermartingale: the certificate buys decrease *only off the sets*,

$$\mathbb{E}[V(X_{t+1}) \mid X_t{=}x] \leq V(x) - \varepsilon, \quad x \notin \mathbb{T} \cup \mathbb{U}.$$

On $\mathbb{T} \cup \mathbb{U}$ nothing constrains the drift — after reaching $\mathbb{T}$ the dynamics can carry V back *up*. So $\mathbb{E}[N_{t+1} \mid \mathcal{F}_t] \leq N_t$ can fail, and the proof rule is simply inapplicable.

Stopping repairs it.

With $M_t = V(X_{t\wedge\tau})$,

- on $\{\tau > t\}$ the state is off $\mathbb{T} \cup \mathbb{U}$: $\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] \leq M_t - \varepsilon$;
- on $\{\tau \leq t\}$ it is constant: $\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] = M_t$.

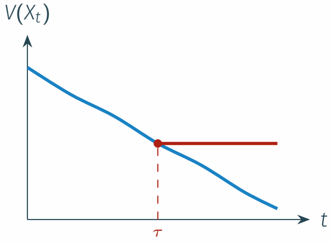
The inequality now holds at *every* t.

Stopping repairs the process exactly where the certificate gives no control.

The stopping rule is a *choice* (and the function does not make it)

The rule we used is a *level set of V*:

$$\tau = \inf\{\, t : V(X_t) \geq \beta \text{ or } X_t \in \mathbb{T} \,\}.$$



Same certificate, two valid processes.

The stopping rule is a *choice* (and the function does not make it)

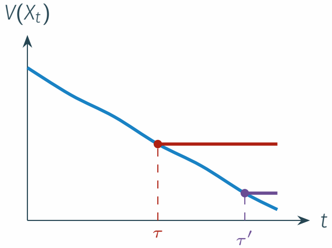
The rule we used is a *level set of V*:

$$\tau = \inf\{\, t : V(X_t) \geq \beta \text{ or } X_t \in \mathbb{T} \,\}.$$

The *same* V admits other valid rules — e.g. stopping on the actual unsafe set,

$$\tau' = \inf\{\, t : X_t \in \mathbb{U} \text{ or } X_t \in \mathbb{T} \,\}.$$

$M_{t\wedge\tau'}$ is equally a supermartingale.



Same certificate, two valid processes.

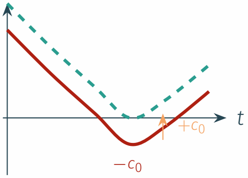
Same problem, same V — V encodes neither choice. That is the crux.

Two lesser freedoms on the function side

These reshape V *deterministically*; they are a milder kind of choice.

A nonnegativity shift

If the network hands you $\inf_x V(x) = -c_0 < 0$, optional stopping (via Doob) wants nonnegativity — so ride $V + c_0 \geq 0$ instead. Same function, shifted.

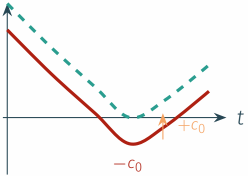


Two lesser freedoms on the function side

These reshape V *deterministically*; they are a milder kind of choice.

A nonnegativity shift

If the network hands you $\inf_x V(x) = -c_0 < 0$, optional stopping (via Doob) wants nonnegativity — so ride $V + c_0 \geq 0$ instead. Same function, shifted.

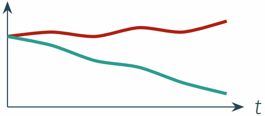


A time compensator

If GV can be *positive*, subtract an increasing A_t that outpaces the deficit:

$$V(X_t) - \left(1 - e^{-t}\right) \quad \text{or} \quad V(X_t) - c\,t.$$

The augmented process is a supermartingale.



Conclusion

The one message

- A supermartingale certificate is a *function*: $V : \mathbb{X} \to \mathbb{R}_{\geq 0}$.
- The object that actually *carries the proof* is a *stochastic process*.
- These are not the same object.

The passage **function** $\longmapsto$ **process** is a real construction — and it is not canonical: one function induces *many* supermartingales.

Open questions

- Compositionality: compose two random systems — can we glue their certificates? (A Fréchet-style combination for $\wedge/\vee$ is a first step.)

Open questions

- Compositionality: compose two random systems — can we glue their certificates? (A Fréchet-style combination for $\wedge/\vee$ is a first step.)
- The standing question: what *is* a good formalization for the **function** $\longmapsto$ **process** construction, with choices and side-conditions included?

The Treachery of Supermartingales: *Ceci n'est pas une supermartingale*

(but it *induces* one — in more ways than one)

Thank you — let's discuss

Related talks in this seminar:

Giacobbe (neural certificates) · Lynch (Markov semigroups) · Corfield, Debauche (Lyapunov)